ARL-TR-8161 • Sep 2017

US Army Research Laboratory

Excluding Noise from Short Krylov Subspace Approximations to the Truncated Singular Value Decomposition (SVD)

by Alex Breuer

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

ARL-TR-8161 ● SEP 2017

US Army Research Laboratory

Excluding Noise from Short Krylov Subspace Approximations to the Truncated Singular Value Decomposition (SVD)

by Alex Breuer

Computational and Information Sciences Directorate, ARL

REPORT DOCUMENTATION PAGE	***Form Approved*** ***OMB No. 0704-0188***

1. REPORT DATE *(DD-MM-YYYY)* September 2017	**2. REPORT TYPE** Technical Report	**3. DATES COVERED (From - To)** October 2015–January 2016

4. TITLE AND SUBTITLE Excluding Noise from Short Krylov Subspace Approximations to the Truncated Singular Value Decomposition (SVD)	**5a. CONTRACT NUMBER**
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S) Alex Breuer	**5d. PROJECT NUMBER**
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) US Army Research Laboratory ATTN: RDRL-CIH-C Aberdeen Proving Ground, MD 21005-5066	**8. PERFORMING ORGANIZATION REPORT NUMBER** ARL-TR-8161
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	**10. SPONSOR/MONITOR'S ACRONYM(S)**
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

primary author's email: `alexander.m.breuer.civ@mail.mil`

14. ABSTRACT

The truncated singular value decomposition (SVD) finds numerous applications, from dimension reduction to matrix regularization to data cleaning. The SVD truncated to rank n produces the rank-n approximation with smallest Frobenius norm error and also separates global structure from local deviations and noise. Fully accurate computation of the truncated SVD is often expensive. Krylov subspace approximations of the truncated SVD may be used to realize substantial computational savings. Approximation of the truncated SVD does not require finding an invariant subspace. The iteration may be terminated after only n iterations. Though these Krylov subspace approximations are close to the truncated SVD with respect to the Frobenius norm, they may not reproduce the important data cleaning qualities of the truncated SVD. We link the presence of noise in Krylov subspaces to the start vector and show necessary and sufficient conditions on the start vector to produce a Krylov subspace that is free of noise up to an arbitrary threshold. These conditions may be used to design stopping criteria for implicitly or explicitly filtering a start vector to effect a noise-free Krylov subspace.

15. SUBJECT TERMS

dimension reduction, Krylov subspaces, truncated SVD, iterative methods, matrix regularization

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** Alex Breuer
a. REPORT Unclassified	**b. ABSTRACT** Unclassified	**c. THIS PAGE** Unclassified	UU	32	**19b. TELEPHONE NUMBER (Include area code)** 410-278-5525

Contents

List of Figures

List of Tables

Acknowledgments

Dr Breuer would like to acknowledge the work of Dr Andrew Lumsdaine (Pacific Northwest National Laboratory) and Dr Geoffrey Sanders (Lawrence Livermore National Laboratory) who assisted in the preparation of this report.

1. Introduction

Truncated singular value decompositions (tSVDs) are used for a variety of tasks in domains ranging from dimension reduction using principal component analysis (PCA),[1] to eigenfaces[2] in machine learning, to latent semantic indexing (LSI)[3] in information retrieval, matrix regularization,[4,5] and sundry methods in signal processing.[6,7] It is the core operation of data analysis techniques that use diagonal matrix factorizations, such as PCA[1,8] and proper orthogonal decomposition.[9] In all cases, one is presented with M points $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_m$ embedded in $\mathbb{R}^N$. These are assembled into a matrix $\mathbf{A} \in \mathbb{R}^{N \times M}$; the goal is to find a reduced-dimension approximation $\hat{\mathbf{A}}$ of $\mathbf{A}$, where $\hat{\mathbf{A}} \in \mathbb{R}^{n \times M}$, with $n \ll N$. With the tSVD, the data are projected into the space spanned by a small subset of singular vectors; these are the n singular vectors that have the n largest singular values. In particular, the tSVD provides 2 key advantages for dimension reduction applications:

- it approximates the data in a lower dimension; thereby reducing storage and processing costs while maintaining important features of the data, and
- it exhibits data cleaning properties by projecting into a space orthogonal to dimensions along which variance is relatively small.

The latter of the above properties—data cleaning—is an important feature of tSVD methods for dimension reduction and data approximation. Partitioning $\mathbb{R}^N$ with the tSVD of $\mathbf{A}$ has been shown to separate global structure of columns of $\mathbf{A}$ from local deviations and noise.[8,10] Global structure is represented by left singular vectors of $\mathbf{A}$ with large singular values, while noise is represented by left singular vectors of $\mathbf{A}$ with small singular values. Important dimension reduction methods that use the tSVD exhibit better performance with reduced dimension data than with the original, high-dimension data; for example, LSI produces reduced dimension approximations that have better precision and recall than is witnessed with the same queries on the original data.[3,11]

A chief drawback of tSVD methods is that they are computationally expensive. This drawback has led several authors to develop approximations to the tSVD that are computationally cheaper. Several methods that substitute a Krylov subspace for a truncated singular vector space have been proposed,[11–15] and they have shown great promise for reducing computational costs while only yielding small differences in the sum-of-squared dimension reduction error of the tSVD. However, analyses of these Krylov subspace methods only have considered the sum-of-squares approximation error difference between the tSVD and a Krylov subspace approximation. The data cleaning properties are not considered.

It is somewhat well known that one cannot generate a sequence of Krylov subspaces that are all orthogonal to the smallest extremal eigenvectors—ones with the smallest eigenvalues, no matter what one does to the start vector. Bounds in Golub et al.[16] show this quantitatively; in fact, if one has a random start vector, then a Krylov subspace used as a tSVD approximation may have a significant overlap with the noise space (we quantitatively define the noise space in Section 1.3 and subspace overlap in Section 2.1) after a small number of iterations. Without filtering the start vector, the data cleaning properties of the Krylov subspace approximation methods are poor. The presence of noise destroys the important data cleaning advantages of low-rank data approximation.

1.1 Summary of Contents

Krylov subspace tSVD approximations[11–15] will almost certainly not remove noise as well as the tSVD when the start vector of the Krylov subspace is random. Even when the start vector is not random, if the start vector is not orthogonal to all noise content, then noise will eventually be present in the Krylov subspace.

It is well known that the principal angles between a Krylov subspace and the noise-space vectors will shrink as one grows a Krylov subspace, no matter how close to orthogonal the start vector is to the noise space. However, this relationship has not been well studied; there is no theory to predict how “noisy” a Krylov subspace approximation of the tSVD will be. Our main result bounds the noise content of Krylov subspaces as dimension reduction projections. We present sufficient conditions that guarantee noise filtering of the Krylov subspace based on the noise content of the initial vector.

These sufficient conditions allow one to design a filtering procedure to produce start vectors that generate Krylov subspaces with bounded noise content. One may use Krylov subspace tSVD approximation methods, and enjoy the noise cleaning properties of the tSVD while preserving the significant computational cost savings of the Krylov subspace matrix approximation methods.[11–15]

1.2 Dimension Reduction and the tSVD

The exact nature of the data approximation and analysis problem assumes that we have data embedded in an N-dimensional space; that is, our data are in the form of vectors $\mathbf{a}_i \in \mathbb{R}^N$ for $i = 1, 2, \ldots, M$. We further assume that all the data $\mathbf{a}_i$ are assembled into a matrix $\mathbf{A} = [\mathbf{a}_1\ \mathbf{a}_2\ \cdots\ \mathbf{a}_M]$. Then, all the methods mentioned in the previous section obtain dimension reduction and data cleaning via the tSVD of $\mathbf{A}$ or some matrix directly derived from $\mathbf{A}$. The chief differences between PCA, LSI, eigenfaces, proper orthogonal decomposition,[9] and the like, lie in the derivation steps applied to $\mathbf{A}$ before the tSVD.

We now proceed to formally explain the SVD and define the tSVD. We assume without any loss of generality that $N \geq M$—otherwise simply replace $\mathbf{A}$ with $\mathbf{A}^\mathsf{T}$. Then, any $N \times M$ matrix $\mathbf{A}$ can be factorized as

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\mathsf{T} \tag{1}$$

with $\mathbf{U} \in \mathbb{R}^{N\times M}$ and $\mathbf{V} \in \mathbb{R}^{M\times M}$ with orthonormal columns, $\mathbf{\Sigma}$ diagonal and having shape $M \times M$. All diagonal elements of Σ are real and nonnegative. Columns of $\mathbf{U}$ and $\mathbf{V}$ are called left and right singular vectors of $\mathbf{A}$, and diagonal elements of Σ are called singular values of $\mathbf{A}$. We write $\sigma_i(\mathbf{A})$ to denote a diagonal element of Σ, and order them such that $\sigma_1(\mathbf{A}) \geq \sigma_2(\mathbf{A}) \geq \cdots \geq \sigma_M(\mathbf{A})$. The tuples $(u_i, v_i, \sigma_i(\mathbf{A}))$ are singular triplets of $\mathbf{A}$. We call a singular vector, value, or triplet leading if i close to 1, and trailing if i is close to M.

Definition 1: tSVD

Let $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\mathsf{T}$ be the SVD of $\mathbf{A}$. Then the tSVD of $\mathbf{A}$ is given by

$$\hat{\mathbf{A}}^{(n)}_{\mathrm{tSVD}} = \mathbf{U}_n\mathbf{\Sigma}_{n,n}\mathbf{V}_n^\mathsf{T} \tag{2}$$

where $\mathbf{U}_n = [\mathbf{u}_1\ \mathbf{u}_2\ \cdots\ \mathbf{u}_n]$, $\mathbf{V}_n = [\mathbf{v}_1\ \mathbf{v}_2\ \cdots\ \mathbf{v}_n]$, and $\mathbf{\Sigma}_{n,n} = \mathrm{diag}(\sigma_1(\mathbf{A}), \sigma_2(\mathbf{A}), \ldots, \sigma_n(\mathbf{A}))$. That is, $\mathbf{U}_n$ and $\mathbf{V}_n$ are composed of the leading n left and right singular vectors, respectively, and $\mathbf{\Sigma}_{n,n}$ is composed of the leading n singular values.

It is clear from the definition of the tSVD that it is a projection of $\mathbf{A}$ through the n-dimensional space colspan$\{\mathbf{U}_n\}$ formed by the span of the leading left singular vectors.

Also, the n-dimensional tSVD is optimal with respect to the Frobenius norm, so $\hat{\mathbf{A}}^{(n)}_{\text{tSVD}} = \arg\min_{\text{rank}(\hat{\mathbf{A}})=n} \|\hat{\mathbf{A}} - \mathbf{A}\|_F$. The tSVD produces the reduction of $\mathbf{A}$ to n-dimensions that is optimal in the least-squares sense aggregated over all $\mathbf{a}_i$.

1.3 Signal and Noise Spaces

We have noted that truncating the SVD also can de-noise the data. We define signal and noise spaces in terms of the SVD.

Definition 2: Signal and Noise Space

Suppose $\mathbf{A}$ is an arbitrary real matrix with SVD $\mathbf{A} = \mathbf{U\Sigma V}^\mathsf{T}$. Let $0 < \tau < 1$ be a nonnegative real number. Then the noise space $\mathcal{U}_{\text{noise}}$ of $\mathbf{A}$ is defined as

$$\mathcal{U}_{\text{noise}} = \text{span}\left\{\mathbf{u}_p, \mathbf{u}_{p+1}, \ldots, \mathbf{u}_M\right\} \tag{3}$$

where p is the smallest natural number that satisfies

$$\sqrt{\frac{\sum_{i=1}^{p} \sigma_i(\mathbf{A})^2}{\sum_{i=1}^{M} \sigma_i(\mathbf{A})^2}} > \tau. \tag{4}$$

The signal space $\mathcal{U}_{\text{signal}}$ of $\mathbf{A}$ is then defined as the complement of $\mathcal{U}_{\text{noise}}$:

$$\mathcal{U}_{\text{signal}} = \mathcal{U}_{\text{noise}}^{\perp}. \tag{5}$$

Remark 1. When the mean of the columns $\mathbf{a}_i$ of $\mathbf{A}$ are zero centered and therefore $\sum_{i=1}^{M} \mathbf{a}_i = 0$, the Gram matrix $\mathbf{AA}^\mathsf{T}$ satisfies $\mathbf{AA}^\mathsf{T} = s\mathbf{C}$, where s is some scalar and $\mathbf{C}$ is the covariance matrix of the sample $\mathbf{a}_i$. In this case, a tSVD of $\mathbf{A}$ is equivalent to projecting out the $N - n$ orthogonal dimensions along which variance is smallest.

Thus, the noise space of $\mathbf{A}$ is defined as the space in which less than τ of the Frobenius norm of $\mathbf{A}$ lies. It is clear that as $\tau \to 1$, p approaches that index of the smallest nonzero singular value. This illustrates that the "noisiest" singular vectors are those with the smallest singular values.

When building a Krylov subspace for low-rank approximation, we want to guarantee that it is (nearly) orthogonal to these noisiest singular vectors for some τ that is close to 1.

1.4 Minimal Krylov Subspaces for Approximation of the tSVD

Though the tSVD has these advantageous properties, it can be expensive to compute, especially if n is on the order of tens or more. A number of authors have proposed Krylov subspaces as a surrogate for the leading singular vector space for dimension reduction tasks.[11–15] Krylov subspaces are defined in terms of square matrices. When $\mathbf{A}$ is not square, Krylov subspaces are typically defined in terms of the Gram matrices $\mathbf{G} = \mathbf{A}\mathbf{A}^{\mathsf{T}}$ or $\mathbf{G}_{\mathsf{T}} = \mathbf{A}^{\mathsf{T}}\mathbf{A}$. These 2 Gram matrices transform the singular value problem into an equivalent eigenvalue problem on $\mathbf{G}$ or $\mathbf{G}_{\mathsf{T}}$; given the SVD $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathsf{T}}$,

$$\mathbf{G} = \mathbf{A}\mathbf{A}^{\mathsf{T}} = \mathbf{U}\mathbf{\Sigma}^2\mathbf{U}^{\mathsf{T}} \tag{6}$$

and

$$\mathbf{G}_{\mathsf{T}} = \mathbf{A}^{\mathsf{T}}\mathbf{A} = \mathbf{V}\mathbf{\Sigma}^2\mathbf{V}^{\mathsf{T}}. \tag{7}$$

Hereafter, when we write λ_i, it is implied that $\lambda_i = \lambda_i(\mathbf{G}) = \lambda_i(\mathbf{G}_{\mathsf{T}}) = \sigma_i(\mathbf{A})^2$

Definition 3: Krylov Subspace

Suppose $\mathbf{G} = \mathbf{A}\mathbf{A}^{\mathsf{T}}$ with shape $N \times N$ and $\mathbf{z}^{(0)} \in \mathbb{R}^N$. Then the nth Krylov subspace is given by

$$\mathcal{K}_n\left(\mathbf{G}, \mathbf{z}^{(0)}\right) = \operatorname{span}\left\{\mathbf{z}^{(0)}, \mathbf{G}\mathbf{z}^{(0)}, \mathbf{G}^2\mathbf{z}^{(0)}, \ldots, \mathbf{G}^{n-1}\mathbf{z}^{(0)}\right\}. \tag{8}$$

We call $\mathbf{z}^{(0)}$ the start vector of the Krylov subspace. Approximation error of a singular vector $\mathbf{u}_i$ of $\mathbf{A}$ depends on the angle $\vartheta(\mathbf{z}^{(0)}, \mathbf{u}_i) = \cos^{-1}\left\langle\mathbf{z}^{(0)}, \mathbf{u}_i\right\rangle / ||\mathbf{z}^{(0)}||||\mathbf{u}_i||$, where $\langle\cdot,\cdot\rangle$ is the inner product, and on the distribution of singular values of $\mathbf{A}$. The closer $\mathbf{z}^{(0)}$ is to $\mathbf{u}_i$—the larger $\cos\vartheta\left(\mathbf{z}^{(0)}, \mathbf{u}_i\right)$—the better the approximation of $\mathbf{u}_i$ in $\mathcal{K}_n\left(\mathbf{G}, \mathbf{z}^{(0)}\right)$. When no beforehand information is available, $\mathbf{z}^{(0)}$ is typically chosen to be random.

Krylov subspace methods have been well proven as iterative SVD solvers.[16,17] One projects $\mathbf{A}$ into the intersection of $\mathcal{K}_n\left(\mathbf{G}, \mathbf{z}^{(0)}\right)$ and $\mathcal{K}_n\left(\mathbf{G}_{\mathsf{T}}, \mathbf{G}\mathbf{z}^{(0)}\right)$, and each iteration reduces approximation errors of extremal singular values. In fact, the tSVD is often computed with a Krylov subspace solver. The difference is that to compute an n-dimensional tSVD, one will likely need to generate a Krylov subspace $\mathcal{K}_k\left(\mathbf{G}, \mathbf{z}^{(0)}\right)$ with $k \gg n$, and/or repeatedly select a better start vector and generate a new Krylov subspace. The tSVD approximation methods[11–15] instead use $k = n$: a minimal Krylov subspace.

Definition 4: Minimal Krylov Subspace

The kth Krylov subspace $\mathcal{K}_k\left(\mathbf{G}, \mathbf{z}^{(0)}\right)$ is said to be minimal for a reduction to n dimensions if $k = n$.

Remark 2. In typical use of Krylov subspaces, one generates a subspace much larger than the solution space that is needed. The wanted solutions are then extracted from the Krylov subspace. For example, if one wants to compute n eigenvalues, one will generate $\mathcal{K}_k\left(\mathbf{G}, \mathbf{z}^{(0)}\right)$ with $k > n$, and often $k \gg n$. When all n eigenvectors are invariant to tolerance, they are extracted from the Krylov subspace, and the problem is projected into the space spanned by those n computed eigenvectors.

1.5 Approximate Eigenvectors and Eigenvalues from Krylov Subspaces

An orthonormal basis $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n$ of approximate eigenvectors of $\mathbf{G}$ may be extracted from a Krylov subspace $\mathcal{K}_n\left(\mathbf{G}, \mathbf{z}^{(0)}\right)$; alternately, these are approximate left singular vectors of $\mathbf{A}$. These vectors are also orthonormal and $\mathbf{G}$-conjugate. We call an approximate eigenvector $\mathbf{z}_i$ from a Krylov subspace a Ritz vector and the value $\theta_i = \mathbf{z}_i^{\mathsf{T}}\mathbf{G}\mathbf{z}_i$ a Ritz value. Any Ritz vector may be expressed as a linear combination of eigenvectors of $\mathbf{G}$.

A Ritz vector $\mathbf{z}_i$ is not necessarily equal to the projection of an eigenvector $\mathbf{Q}_n\mathbf{Q}_n^{\mathsf{T}}\hat{\mathbf{u}}_i\mathbf{u}_i$ (where $\mathbf{Q}_n$ is an orthonormal basis for $\mathcal{K}_n\left(\mathbf{G}, \mathbf{z}^{(0)}\right)$) through the Krylov subspace. That is, we may—and very likely will—have $\mathbf{z}_i \neq \mathbf{Q}_n\mathbf{Q}_n^{\mathsf{T}}\mathbf{u}_i$ for all i. This is because $\mathbf{z}_i$ is defined as

$$\mathbf{z}_i = \min_{\dim(C)=i-1} \arg\max_{\mathbf{x}\perp C} \frac{\mathbf{x}^{\mathsf{T}}\mathbf{G}\mathbf{x}}{||\mathbf{x}||^2}, \tag{9}$$

while $\mathbf{Q}_n\mathbf{Q}_n^{\mathsf{T}}\mathbf{u}_i$ is given by

$$\mathbf{Q}_n\mathbf{Q}_n^{\mathsf{T}}\mathbf{u}_i = \arg\min_{\mathbf{x}\in\mathcal{K}_n\left(\mathbf{G},\mathbf{z}^{(0)}\right)} ||\mathbf{u}_i - \mathbf{x}||. \tag{10}$$

In a minimal Krylov subspace, it is almost certain that there are many Ritz vectors that are not $\mathbf{G}$-invariant; that is, the residual $||\mathbf{G}\mathbf{v}_i - \theta_i\mathbf{v}_i||$ is greater than machine epsilon. Ritz vectors from Krylov subspaces are defined in terms of a polynomial $q(x)$ whose roots are the Ritz values.[18] Thus $q(\theta_i) = 0$. Noise content of the Ritz vector depends on the ratio $\sum_{j=p}^{N} q(\lambda_j)^2 \left\langle \mathbf{z}^{(0)}, \mathbf{u}_j \right\rangle^2 / \sum_{j=1}^{N} q(\lambda_j)^2 \left\langle \mathbf{z}^{(0)}, \mathbf{u}_j \right\rangle^2$, where p is from Definition 2. Whenever $q(\lambda_j)^2 \left\langle \mathbf{z}^{(0)}\mathbf{u}_j \right\rangle^2$ is not tiny for $j \geq p$, then the Ritz vector may not be orthogonal to the noise space.

1.6 A Motivating Example

Constructing a linear classifier is a task that can motivate our discussion. A simple way to construct a linear classifier is to solve $\mathbf{Gx} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$, where $\mathbf{G}$ is the covariance matrix and $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are the means of the 2 classes. This problem is ill-posed when $\mathbf{G}$ has small singular values (a nontrivial noise space), and we want a classifier that is orthogonal to those dimensions along which variance is small.

The example is a matrix regularization problem. In matrix regularization problems,[19] one has a matrix $\mathbf{G}$ that has small singular values and one seeks a solution that minimizes $||\mathbf{Gx} - \mathbf{b}||$ but is also orthogonal to the singular vector space corresponding to the small singular vectors of $\mathbf{G}$. The small singular values of $\mathbf{G}$ make minimization of $||\mathbf{Gx} - \mathbf{b}||$ ill-posed, as small perturbations to $\mathbf{b}$ may result in a large perturbation of $\mathbf{x}$. One instead minimizes a regularized problem such as $||\mathbf{Gx} - \mathbf{b}|| + \eta||\mathbf{x}||$, where η is a user-chosen regularization parameter picked to avoid small singular vectors of $\mathbf{G}$. Using a tSVD can be as effective as directly minimizing $||\mathbf{Gx} - \mathbf{b}|| + \eta||\mathbf{x}||$ for optimal η.[4,5,20]

If one substitutes a minimal Krylov subspace $\mathcal{K}_n\left(\mathbf{G}, \mathbf{z}^{(0)}\right)$ with a random $\mathbf{z}^{(0)}$ for the truncated singular vector space, then the influence of small singular values is difficult to control either directly, as with the tSVD, or indirectly, as in explicit minimization of $||\mathbf{Gx} - \mathbf{b}|| + \eta||\mathbf{x}||$. An $\mathbf{x}$ computed with an $\hat{\mathbf{A}}^{(n)}$ from a minimal Krylov subspace may have a large $||\mathbf{x}||$, which would have been avoided with even a small η.

Example 1. Let $\mathbf{G}$ be a $2,000 \times 2,000$ diagonal matrix defined as

$$\mathbf{G} = \mathrm{diag}(1, 1/2, 1/3, 1/4, \ldots, 1/1999, 10^{-17}). \tag{11}$$

Since $\mathbf{G}$ is diagonal, its diagonal entries are its eigenvalues. Moreover, since all its eigenvalues are nonnegative, its spectral decomposition and SVD coincide. The spectrum of $\mathbf{G}$ is shown in Fig. 1.

Set $\mathbf{z}^{(0)} = 1/\sqrt{2000}\sum_{i=1}^{2000}\mathbf{u}_i$. We have $\cos\vartheta\left(\mathbf{z}^{(0)}, \mathbf{u}_i\right) = 1/\sqrt{2000}$ for all eigenvectors $\mathbf{u}_i$. We compute an orthonormal basis $\mathbf{Q}_n$ for $\mathcal{K}_n\left(\mathbf{G}, \mathbf{z}^{(0)}\right)$ and the $\hat{\mathbf{G}} = \mathbf{Q}_n^\mathsf{T}\mathbf{GQ}$ for $1 \leq n \leq 20$. We generate a random $\mathbf{b}$ and solve the least squares problem $\mathbf{x} = \arg\min_{\mathbf{y}} ||\mathbf{T}_{n,n}\mathbf{y} - \mathbf{b}||$.

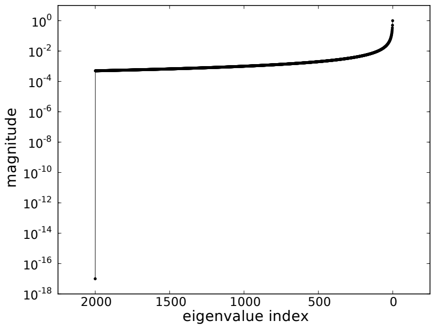

Fig. 1 Spectrum of G as defined in Eq. 11

We compute the Frobenius norm of $\hat{\mathbf{G}}^{-1}$ where $\hat{\mathbf{G}} = \mathbf{Q}_n^{\mathsf{T}}\mathbf{G}\mathbf{Q}$ for the Krylov subspace solution or $\hat{\mathbf{G}} = \mathbf{U}_n\mathbf{\Sigma}_{n,n}\mathbf{U}_n^{\mathsf{T}}$ for the tSVD solution; the values are shown in Fig. 2. Values of $||\mathbf{x}||$ are also shown for both the tSVD and minimal Krylov subspace $\mathbf{x}$. Figures 1 and 2 show that substituting a minimal Krylov subspace for a truncated singular vector space for matrix regularization produces poorer regularization results. The Frobenius norms of the $\hat{\mathbf{G}}^{-1}$ from the minimal Krylov subspace are at least 10 times larger than the tSVD $\hat{\mathbf{G}}^{-1}$, and the $||\mathbf{x}||$ are at least 100 times larger for the minimal Krylov subspace.

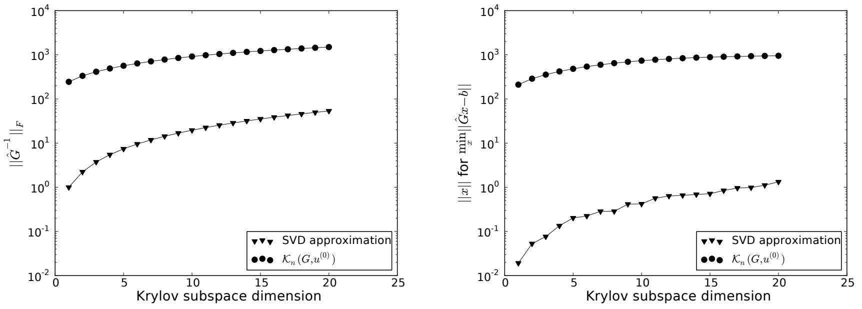

Fig. 2 Frobenius norms of $\hat{\mathbf{G}}^{-1}$, where $\hat{\mathbf{G}}$ is G restricted to $\mathcal{K}_n\left(\mathbf{G}, \mathbf{z}^{(0)}\right)$ or restricted to the truncated singular vector space span $\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$ **(left). Norms $||x||$ where x is a solution to the least squares problem $||\hat{\mathbf{G}}\mathbf{x} - \mathbf{b}||$ and where $\hat{\mathbf{G}} = \mathbf{Q}_n^{\mathsf{T}}\mathbf{G}\mathbf{Q}_n$ where $\mathbf{Q}_n$ is an orthonormal basis for $\mathcal{K}_n\left(\mathbf{G}, \mathbf{z}^{(0)}\right)$ or $\hat{\mathbf{G}} = \mathbf{U}_n\mathbf{\Sigma}_{n,n}^{\mathsf{T}}\mathbf{U}_n^{\mathsf{T}}$ (right). Large values indicate greater influence of small singular vectors in x and more sensitivity to small perturbations in b.**

2. Corruption of Subspaces with Noise

When the basis vectors of the Krylov subspace are not orthogonal to the noise space, then the Krylov subspace has been "corrupted" by noise. The proceeding analysis requires a measurement of how much a subspace is corrupted with noise. It is evident that measurement of the corruption of a space $\mathcal{S}$ by noise is equivalent to measuring the norms of images of noise space basis vectors u_j projected into S. We use the principal angles between spaces to formalize this concept.

2.1 Principal Angles for Quantifying Subspace Overlap

Much of the proceeding analysis considers the principal angles between spaces (see Zhu and Knyazev[21][Definition 2.1 and Theorem 2.1]). We use these to quantify the overlap between 2 subspaces of $\mathbb{R}^N$. In the context of our analysis of minimal Krylov subspaces for approximating the tSVD, we would like to have the overlap between $\mathcal{K}_n\left(\mathbf{G}, \mathbf{z}^{(0)}\right)$ and $\mathcal{U}_{\text{noise}}$ as small as possible.

Definition 5: Principal Angles
Let $\mathbf{X}$ and $\mathbf{Y}$ be matrices in $\mathbb{R}^N$ with orthonormal columns. Then the principal angles between colspan $\{\mathbf{X}\}$ and colspan $\{\mathbf{Y}\}$ are defined as

$$\vartheta(\mathbf{X}, \mathbf{Y}) = [\cos^{-1}(\sigma_1(\mathbf{X}^\mathsf{T}\mathbf{Y})) \ \cos^{-1}(\sigma_2(\mathbf{X}^\mathsf{T}\mathbf{Y})) \ \cdots] = \cos^{-1}(\sigma(\mathbf{X}^\mathsf{T}\mathbf{Y})). \tag{12}$$

When either $\mathbf{X}$ or $\mathbf{Y}$ has only one column, then there is only one principal angle. There is also a close relationship among the principal angles $\vartheta(\mathbf{X}, \mathbf{Y})$, the Frobenius norm $\|\mathbf{X}^\mathsf{T}\mathbf{Y}\|_F$, and the spectral norm $\|\mathbf{X}^\mathsf{T}\mathbf{Y}\|_2$. The spectral norm is $\|\mathbf{X}^\mathsf{T}\mathbf{Y}\|_2 = \sqrt{\cos\vartheta_1(\mathbf{X}, \mathbf{Y})^2} = \sigma_1(\mathbf{X}^\mathsf{T}\mathbf{Y})$ and the Frobenius norm is $\|\mathbf{X}^\mathsf{T}\mathbf{Y}\|_F = \sqrt{\sum_{i=1}^k \cos\vartheta_i(\mathbf{X}, \mathbf{Y})^2} = \sqrt{\sum_{i=1}^k \sigma_i(\mathbf{X}^\mathsf{T}\mathbf{Y})^2}$ where $\mathbf{X}^\mathsf{T}\mathbf{Y}$ has k nonzero singular values.

Clearly, when colspan $\{\mathbf{X}\} \perp$ colspan $\{\mathbf{Y}\}$, then $\vartheta_i(\mathbf{X}, \mathbf{Y}) = \pi/2$ for all valid i, and $\|\mathbf{X}^\mathsf{T}\mathbf{Y}\|_F = \|\mathbf{X}^\mathsf{T}\mathbf{Y}\|_2 = 0$. The closer all principal angles are to $\pi/2$, the smaller the overlap between colspan $\{\mathbf{X}\}$ and colspan $\{\mathbf{Y}\}$.

2.2 A Measure of Corruption: ρ-Free of Noise

Principal angles and the closely related matrix norms $\|\cdot\|_F$ and $\|\cdot\|_2$ lead naturally to a measure of the overlap between the noise space $\mathcal{U}_{\text{noise}}$ and some subspace S with orthonormal basis W: ρ-free of noise.

Definition 6: ρ-Free of Noise

Let $\mathcal{S} \subset \mathbb{R}^N$ be some subspace, let $\mathbf{W}$ be an orthonormal basis for $\mathcal{S}$, let $\mathcal{U}_{\text{noise}} \subset \mathbb{R}^N$ be the noise space of $\mathbf{G}$, and $\mathbf{U}_{\text{noise}}$ be an orthonormal basis for $\mathcal{U}_{\text{noise}}$. Pick some nonnegative real ρ. Then $\mathcal{S}$ is ρ-free of noise if

$$\cos \vartheta_1 \left(\mathbf{U}_{\text{noise}}, \mathbf{W}\right) \leq \rho. \tag{13}$$

An equivalent condition to Eq. 13 is that $\|\mathbf{U}_{\text{noise}}^{\mathsf{T}}\mathbf{W}\|_2 \leq \rho$. Since $\|\mathbf{U}_{\text{noise}}^{\mathsf{T}}\mathbf{W}\|_2 \leq \|\mathbf{U}_{\text{noise}}^{\mathsf{T}}\mathbf{W}\|_F$, $\|\mathbf{U}_{\text{noise}}^{\mathsf{T}}\mathbf{W}\|_F \leq \rho$ also implies that $\mathcal{S}$ is ρ-free of noise.

3. Two Sufficient Conditions on $\mathbf{z}^{(0)}$ for $\mathcal{K}_n\left(\mathbf{G}, \mathbf{z}^{(0)}\right)$ to Be ρ-Free of Noise

We are now ready to present our main result. We develop 2 criteria that guarantee that the nth Krylov subspace is ρ-free of noise.

3.1 A Basic Sufficient Condition on $\mathbf{z}^{(0)}$ for an Uncorrupted Subspace

Our basic sufficient condition comes from Corollary 1, which depends on Lemma 1 and Theorem 1. First, we use the Lanczos recurrence (see Saad,[22] Section 3.2) to bound the cosine $\cos \vartheta \left(\mathbf{q}_{n+1}, \mathbf{u}_i\right)$ in Lemma 1, where $\mathbf{q}_{n+1}$ is a basis vector generated by the Lanczos algorithm.[22] This result then leads to a recurrence relation that bounds the image $\hat{\mathbf{u}}_i^{(n)}$ of the eigenvector $\mathbf{u}_i$ projected into $\mathcal{K}_n\left(\mathbf{G}, \mathbf{z}^{(0)}\right)$ in Theorem 1; the sufficient condition in Corollary 1 on $\mathbf{z}^{(0)}$ follows from that.

We begin with bounding the cosine $\cos \vartheta \left(\mathbf{q}_{n+1}, \mathbf{u}_i\right)$.

Lemma 1

Let $\mathbf{q}_{n+1}$ be the $n+1$th basis vector generated by the Lanczos algorithm acting on a Gram matrix $\mathbf{G}$ and $\mathbf{z}^{(0)}$. Let $\mathbf{Q}_n$ be an orthonormal basis for $\mathcal{K}_n\left(\mathbf{G}, \mathbf{z}^{(0)}\right)$. Let $\mathbf{T}_{n,n} = \mathbf{Q}_n^{\mathsf{T}}\mathbf{G}\mathbf{Q}_n$ be the restriction of $\mathbf{G}$ to $\mathcal{K}_n\left(\mathbf{G}, \mathbf{z}^{(0)}\right)$, and $\hat{\mathbf{u}}_i^{(n)} = \mathbf{Q}_n^{\mathsf{T}}\mathbf{u}_i$. Note that $\mathbf{T}_{n,n}$ is a tridiagonal matrix.[22] Let β_{n+1} be the norm of the residual of the Lanczos algorithm after the nth step. Order the singular values of a matrix $\mathbf{A}$ as $\sigma_1(\mathbf{A}) \geq \sigma_2(\mathbf{A}) \geq \cdots \geq \sigma_n(\mathbf{A})$, and the eigenvalues of $\mathbf{G}$ as $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N$. Then

$\cos\vartheta\left(\mathbf{u}_i, \mathbf{q}_{n+1}\right)$ obeys

$$\cos\vartheta\left(\mathbf{u}_i, \mathbf{q}_{n+1}\right) \leq \frac{\|\hat{\mathbf{u}}_i^{(n)}\|\sigma_1(\mathbf{T}_{n,n} - \mathbf{I}\lambda_i)}{\beta_{n+1}}.$$

Proof. Let $\mathbf{Q}_n$ be the orthonormal basis generated for $\mathcal{K}_n\left(\mathbf{G}, \mathbf{z}^{(0)}\right)$ by the Lanczos algorithm after n steps. Then we have

$$\mathbf{G}\mathbf{Q}_n = \mathbf{Q}_n\mathbf{T}_{n,n} + \mathbf{r}_n\mathbf{e}_n^{\mathsf{T}}.$$

Write the inner product on $\mathbb{R}^N$ as $\langle\cdot,\cdot\rangle$. Left-multiplying both sides by $\mathbf{u}_i^{\mathsf{T}}$ gives

$$\begin{aligned} \mathbf{u}_i^{\mathsf{T}}\mathbf{G}\mathbf{Q}_n &= \mathbf{u}_i^{\mathsf{T}}\mathbf{Q}_n\mathbf{T}_{n,n} + \mathbf{u}_i^{\mathsf{T}}\mathbf{r}_n\mathbf{e}_n^{\mathsf{T}} \\ \mathbf{u}_i^{\mathsf{T}}\mathbf{G}\mathbf{Q}_n - \mathbf{u}_i^{\mathsf{T}}\mathbf{Q}_n\mathbf{T}_{n,n} &= \mathbf{u}_i^{\mathsf{T}}\mathbf{r}_n\mathbf{e}_n^{\mathsf{T}} \end{aligned}$$

and

$$\beta_{n+1}\langle\mathbf{u}_i, \mathbf{q}_{n+1}\rangle\mathbf{e}_n^{\mathsf{T}} = \mathbf{u}_i^{\mathsf{T}}\mathbf{G}\mathbf{Q}_n - \mathbf{u}_i^{\mathsf{T}}\mathbf{Q}_n\mathbf{T}_{n,n}$$

as both $\mathbf{q}_{n+1}\beta_{n+1} = \mathbf{r}_n$. Then

$$\beta_{n+1}|\langle\mathbf{u}_i, \mathbf{q}_{n+1}\rangle| = \beta_{n+1}\|\langle\mathbf{u}_i, \mathbf{q}_{n+1}\rangle\mathbf{e}_n^{\mathsf{T}}\| = \|\mathbf{u}_i^{\mathsf{T}}\mathbf{G}\mathbf{Q}_n - \mathbf{u}_i^{\mathsf{T}}\mathbf{Q}_n\mathbf{T}_{n,n}\|$$

and

$$|\langle\mathbf{u}_i, \mathbf{q}_{n+1}\rangle| = \frac{\|\mathbf{u}_i^{\mathsf{T}}\mathbf{G}\mathbf{Q}_n - \mathbf{u}_i^{\mathsf{T}}\mathbf{Q}_n\mathbf{T}_{n,n}\|}{\beta_{n+1}} = \frac{\|\lambda_i\hat{\mathbf{u}}_i^{(n)T} - \hat{\mathbf{u}}_i^{(n)T}\mathbf{T}_{n,n}\|}{\beta_{n+1}}$$

as $\beta_{n+1} \geq 0$. Noting that $\cos\vartheta\left(\mathbf{u}_i, \mathbf{q}_{n+1}\right) = |\langle\mathbf{u}_i, \mathbf{q}_{n+1}\rangle|$ when both vectors are unit-length and

$$\|\hat{\mathbf{u}}_i^{(n)T}(\mathbf{I}\lambda_i - \mathbf{T}_{n,n})\| \leq \|\hat{\mathbf{u}}_i^{(n)}\|\sigma_1(\mathbf{T}_{n,n} - \mathbf{I}\lambda_i) \tag{14}$$

completes the proof. □

Remark 3. Eq. 14 may also be bounded from below as

$$\|\hat{\mathbf{u}}_i^{(n)T}(\mathbf{I}\lambda_i - \mathbf{T}_{n,n})\| \geq \|\hat{\mathbf{u}}_i^{(n)}\|\sigma_n(\mathbf{T}_{n,n} - \mathbf{I}\lambda_i).$$

One could use this result to obtain a different necessary condition for a noise-free Krylov subspace.

We now use the result of Lemma 1 to bound $||\hat{\mathbf{u}}_i^{(n)}||$. This will result in a basic sufficient condition.

Theorem 1

Let $\mathbf{G}$, $\mathbf{z}^{(0)}$, the λ_i, and $\mathbf{T}_{n,n}$ be defined as in Lemma 1. Let β_k be any of the sub- and super-diagonal values of $\mathbf{T}_{n,n}$, and let $\underline{\beta} \leq \beta_k$ for $k = 1, 2, \ldots, n$. Let $\hat{\mathbf{u}}_i^{(n)}$ be the projection of eigenvector $\mathbf{u}_i$ into $\mathcal{K}_n\left(\mathbf{G}, \mathbf{z}^{(0)}\right)$. Suppose that $||\hat{\mathbf{u}}_i^{(0)}|| \leq \varepsilon$ and $\lambda_i \leq \theta_1$, where θ_1 is the principal eigenvalue of $\mathbf{T}_{n,n}$. Then the norm of $\hat{\mathbf{u}}_i^{(n)}$ is bounded as

$$||\hat{\mathbf{u}}_i^{(n)}||^2 \leq \varepsilon^2 \left(\frac{(\lambda_1 - \lambda_i)^2}{\underline{\beta}^2} + 1 \right)^{n-1}.$$

Proof. Let θ_1 be the principal eigenvalue of $\mathbf{T}_{n,n}$. We have $0 \leq \theta_1 \leq \lambda_1$—$\mathbf{G}$ is a Gram matrix and $0 \leq \lambda_N \leq \theta_1$—and it is assumed that $\lambda_i \leq \theta_1$. Then we get

$$\sigma_1(\mathbf{T}_{k,k} - \mathbf{I}\lambda_i) \leq \lambda_1 - \lambda_i.$$

Applying Lemma 1 gives

$$\begin{aligned} \cos\vartheta(\mathbf{u}_i, \mathbf{q}_{n+1}) &\leq \frac{||\hat{\mathbf{u}}_i^{(n)}||(\lambda_1 - \lambda_i)}{\beta_{n+1}} \\ &\leq \frac{||\hat{\mathbf{u}}_i^{(n)}||(\lambda_1 - \lambda_i)}{\underline{\beta}}. \end{aligned}$$

Then we can express $||\hat{\mathbf{u}}_i^{(n)}||$ recursively, as

$$||\hat{\mathbf{u}}_i^{(n)}||^2 = \cos\vartheta(\mathbf{u}_i, \mathbf{q}_n)^2 + ||\hat{\mathbf{u}}_i^{(n-1)}||^2.$$

So

$$\begin{aligned} ||\hat{\mathbf{u}}_i^{(n)}||^2 &\leq \frac{||\hat{\mathbf{u}}_i^{(n-1)}||^2(\lambda_1 - \lambda_i)^2}{\underline{\beta}^2} + ||\hat{\mathbf{u}}_i^{(n-1)}||^2 \\ &\leq ||\hat{\mathbf{u}}_i^{(n-1)}||^2 \left(\frac{(\lambda_1 - \lambda_i)^2}{\underline{\beta}^2} + 1 \right). \end{aligned}$$

The closed form for this series is

$$||\hat{\mathbf{u}}^{(n)}||^2 \leq ||\hat{\mathbf{u}}_i^{(0)}||^2 \left(\frac{(\lambda_1 - \lambda_i)^2}{\underline{\beta}^2} + 1 \right)^{n-1}.$$

Since all noise eigenvectors have $\cos\vartheta\left(\mathbf{u}_i, \mathbf{z}^{(0)}\right) \leq \varepsilon$,

$$\|\hat{\mathbf{u}}_i^{(n)}\|^2 \leq \varepsilon^2 \left(\frac{(\lambda_1 - \lambda_i)^2}{\underline{\beta}^2} + 1\right)^{n-1}$$

for any noise eigenvector $\mathbf{u}_i$. □

This result immediately leads to a sufficient condition on $\mathbf{z}^{(0)}$ for $\mathcal{K}_n\left(\mathbf{G}, \mathbf{z}^{(0)}\right)$ to be ρ-free of noise.

Corollary 1

Let $\mathbf{G}$ and $\mathbf{z}^{(0)}$ be given where eigenvalues of $\mathbf{G}$ are $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N$, and let $N - p$ be the dimension of the noise space $\mathcal{U}_{\text{noise}}$. Suppose that $\|\hat{\mathbf{u}}_i^{(n)}\| \leq \varepsilon$ for any noise eigenvector $\mathbf{u}_i$. Then $\mathcal{K}_n\left(\mathbf{G}, \mathbf{z}^{(0)}\right)$ is ρ-free of noise if

$$\varepsilon^2 (N - p)\left(\frac{(\lambda_1 - \lambda_N)^2}{\underline{\beta}^2} + 1\right)^{n-1} \leq \rho^2. \tag{15}$$

This is due to an application of Theorem 1 to bound the quantity $\|\hat{\mathbf{u}}_N^{(n)}\|$, and noting that $\|\mathbf{u}_M^{\mathsf{T}}\mathbf{Q}_n\| \leq \|\mathbf{U}_{\text{noise}}^{\mathsf{T}}\mathbf{Q}\|_2$. This application of Theorem 1 is always possible, since λ_N is always less than or equal to any eigenvalue of $\mathbf{T}_{n,n}$—the assumption that $\lambda_N \leq \theta_1$ is always valid.

Remark 4. Corollary 1 uses a lower bound $\underline{\beta}$ on the Lanczos residuals β_j. This value is not known a priori, but it was conjectured g[23] that no β_j becomes negligible. The cases for which β_j does attain a small value is when the Krylov subspace is invariant or nearly so. We have observed that the median eigenvalue is often a reasonable lower bound on the β_j from the Lanczos algorithm.

We now proceed with an example in which the median eigenvalue is a reasonable lower bound for $\underline{\beta}$.

Example 2. We continue with the matrix $\mathbf{G}$ defined in Example 1. Set $\rho = 0.001$. Since $\mathcal{U}_{\text{noise}}$ is defined by the trailing 100 eigenvectors, $N - p = 100$. We transform Eq. 15 to a condition on ε:

$$\varepsilon \leq \sqrt{\frac{\rho^2}{(N - p)\left(\frac{(\lambda_1 - \lambda_N)^2}{\underline{\beta}^2} + 1\right)^{n-1}}}. \tag{16}$$

Approved for public release; distribution is unlimited.

We supposed that $\underline{\beta}$ is greater than the median eigenvalue λ_{1000}. We then compute the maximum $\varepsilon_{\max}$ that will satisfy Eq. 16 for $1 \leq n \leq 10$.

To show the pessimism of these $\varepsilon_{\max}$, we also used line search on a posteriori measured values of $\cos(\mathcal{U}_{\text{noise}}, \mathbf{Q}_n)$, where $\mathbf{Q}_n$ is a basis for $\mathcal{K}_n\left(\mathbf{G}, \mathbf{z}^{(0)}\right)$, to find a value $\varepsilon_{\text{computed}}$: a maximum observed value where $\cos\vartheta\left(\mathbf{u}_j, \mathbf{z}^{(0)}\right) \leq \varepsilon_{\text{computed}}$ implies $\mathcal{K}_n\left(\mathbf{G}, \mathbf{z}^{(0)}\right)$ is ρ-free of noise and where $\mathbf{u}_j$ is a noise eigenvector. We computed $\varepsilon_{\text{computed}}$ to 15 significant digits. The values for $\varepsilon_{\max}$ and $\varepsilon_{\text{computed}}$ are shown in the left-hand plot of Fig. 3.

The purpose of this example is to verify the values of $\varepsilon_{\max}$ and $\varepsilon_{\text{computed}}$. We computed bases $\mathbf{Q}_n$ for $\mathcal{K}_n\left(\mathbf{G}, \mathbf{z}^{(0)}\right)$ for $1 \leq n \leq 10$, and used the computation to verify that $\underline{\beta} \geq \lambda_{1000}$. We set

$$\mathbf{z}^{(0)} = \sqrt{\frac{1-(100\varepsilon^2)}{1900}} \sum_{i=1}^{1900} \mathbf{u}_i + \varepsilon \sum_{i=1901}^{2000} \mathbf{u}_i, \tag{17}$$

where ε is either $\varepsilon_{\max}$—defined by equality in Eq. 16, or ε is $\varepsilon_{\text{computed}}$. We then computed an orthonormal basis for $\mathcal{K}_n\left(\mathbf{G}, \mathbf{z}^{(0)}\right)$ and measured $\|\mathbf{U}_{\text{noise}}^{\mathsf{T}}\mathbf{Q}_n\|$. The $\cos\vartheta\left(\mathbf{z}^{(0)}, \mathbf{u}_i\right) = \varepsilon$ for any noise eigenvector $\mathbf{u}_i$. The values of $\|\mathbf{U}_{\text{noise}}^{\mathsf{T}}\mathbf{Q}_n\|_2$ for both $\varepsilon_{\max}$ and $\varepsilon_{\text{computed}}$ are shown in the right plot of Fig. 3.

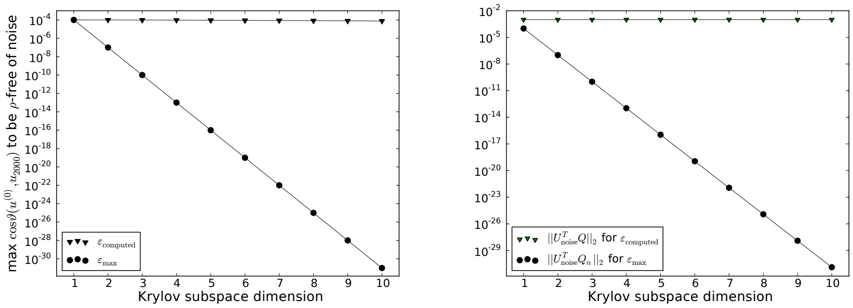

Fig. 3 Values of $\varepsilon_{\max}$ and $\varepsilon_{\text{computed}}$ (left). Values of $\|\mathbf{U}_{\text{noise}}^{\mathsf{T}}\mathbf{Q}_n\|_2$ where $\mathbf{Q}_n$ is a basis for $\mathcal{K}_n\left(\mathbf{G}, \mathbf{z}^{(0)}\right)$ and $\mathbf{z}^{(0)}$ is defined as in Eq. 17 for either $\varepsilon_{\max}$ or $\varepsilon_{\text{computed}}$ (right). The small values of $\|\mathbf{U}_{\text{noise}}^{\mathsf{T}}\mathbf{Q}_n\|_2$ for $\varepsilon_{\max}$ indicate that Corollary 1 is pessimistic for this example.

We notice that Corollary 1 is pessimistic. Satisfying Eq. 15 may be difficult, since $((\lambda_1 - \lambda_N)^2/\underline{\beta}^2)^n$ may grow rapidly. Even for small values of n, producing a sufficiently small $\cos\vartheta\left(\mathbf{z}^{(0)}, \mathbf{u}_j\right)$ may be impossible in finite precision. Even when Eq. 15

is not satisfied, it may be the case that the Krylov subspace is ρ-free of noise. Therefore, we present a tighter a posteriori sufficient condition that uses information from the Krylov subspace.

3.2 A Sharper Sufficient Condition on $\mathbf{z}^{(0)}$

We notice that the basic sufficient condition is pessimistic, and not practical for n greater than 10 or so. Our second sufficient condition gives us extra sharpness to extend a sufficient condition for larger n. Lemma 2, about the polynomial that defines Ritz vectors, directly gives another sufficient condition, which we illustrate in Example 3. A nice feature of the condition that comes from Lemma 2 is that the quantities can be computed from byproducts of the Lanczos algorithm, as is noted in Remark 5.

Lemma 2

Let the positive semidefinite matrix $\mathbf{G}$ and vector $\mathbf{z}^{(0)}$ define the Krylov subspace $\mathcal{K}_n\left(\mathbf{G}, \mathbf{z}^{(0)}\right)$, and let $\theta_1 \geq \theta_2 \geq \cdots \geq \theta_n$ be the Ritz values of the restriction of $\mathbf{G}$ to $\mathcal{K}_n\left(\mathbf{G}, \mathbf{z}^{(0)}\right)$, with unit-length Ritz vectors $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n$. Define the polynomial $q_j(x)$ as

$$q_j(x) = \prod_{\substack{t=1 \\ k \neq j}}^{n} (x - \theta_k)$$

and set $c_i = \left\langle \mathbf{z}^{(0)}, \mathbf{u}_i \right\rangle$, where $\langle \cdot, \cdot \rangle$ is the inner product on $\mathbb{R}^N$. Suppose $p \leq i \leq N$. Then the magnitude of the cosine $\cos \vartheta\left(\mathbf{u}_i, \mathbf{z}_j\right)$ is bounded as

$$\cos \vartheta\left(\mathbf{u}_i, \mathbf{z}_j\right) \leq \max_{0 \leq x \leq \lambda_p} \frac{|q_j(x) c_i|}{||q_j(\mathbf{G})\mathbf{z}^{(0)}||}.$$

Proof. The polynomial

$$q_j(x) = \prod_{\substack{t=1 \\ k \neq j}}^{n} (x - \theta_k)$$

Approved for public release; distribution is unlimited.

gives the j^{th} Ritz vector of $\mathcal{K}_n\left(\mathbf{G}, \mathbf{z}^{(0)}\right)$;[18] $\left(\mathbf{z}_j = \mathbf{q}_j(\mathbf{G})\mathbf{z}^{(0)}/\|\mathbf{q}_j(\mathbf{G})\mathbf{z}^{(0)}\|\right)$. Therefore,

$$\cos\vartheta\left(\mathbf{u}_i, \mathbf{z}_j\right) = \frac{|q_j(\lambda_i)c_i|}{\|q_j(\mathbf{G})\mathbf{z}^{(0)}\|}.$$

Since $i \leq p$ and $\mathbf{G}$ is positive semidefinite, $0 \leq \lambda_i \leq \lambda_p$ and $|q_j(\lambda_i)| \leq \max_{0\leq x\leq\lambda_p} |q(x)|$. Then

$$\cos\vartheta\left(\mathbf{u}_i, \mathbf{z}_j\right) = \max_{0\leq x\leq\lambda_p} \frac{|q_j(x)c_i|}{\|q_j(\mathbf{G})\mathbf{z}^{(0)}\|},$$

which completes the proof. □

Remark 5. The polynomial $q_j(x)$ and the norm $\|q_j(\mathbf{G})\mathbf{z}^{(0)}\|$ can be computed a posteriori inexpensively as a by-product of a standard Krylov subspace algorithm, such as the Lanczos algorithm. However, c_i is typically unknown, but can be bounded if $\mathbf{z}^{(0)}$ has known properties. For example, if $\mathbf{z}^{(0)}$ is a random vector, then $|c_i|$ may be probabilistically bounded with enough tightness as to result in tight bounds for $\cos\vartheta\left(\mathbf{u}_i, \mathbf{z}\right)$. For our example case, where eigenvectors are all standard basis vectors (in Dettman,[24] p. 111) and $\langle\mathbf{u}_i, \mathbf{x}\rangle = x_i$ for x_i is the ith entry of $\mathbf{x}$, this is straightforward.

Example 3. Let $\mathbf{G}$ be defined as in Example 1. Set $\mathbf{z}^{(0)}$ to be a random vector with entries drawn from the normal distribution $\mathcal{N}(0, 1)$. We compute $\mathcal{K}_6\left(\mathbf{G}, \mathbf{z}^{(0)}\right)$ and compute the Ritz values and $q_j(x)$ for $1 \leq j \leq 6$. Each $q_j(x)$ is a quintic polynomial; their derivatives give their maxima between Ritz values θ_i, are quartic, and can be solved analytically. We computed $\theta_6 > \lambda_N$ and $\max_{0\leq x\leq\lambda_p} |q_j(x)| \leq \max_{0\leq x\leq\theta_6} |q_j(x)|$, so it is sufficient to consider maxima of $|q_j(x)|$ for $0 \leq x \leq \theta_6$. For all $q_j(x)$, $|q_j(0)| = \max_{0\leq x\leq\lambda_p} |q_j(x)|$. We now proceed to bound c_i.

Since $\mathbf{z}^{(0)}$ is random, zero-centered, and normally distributed, we may place an upper bound on $\cos\vartheta\left(\mathbf{z}^{(0)}, \mathbf{u}_j\right)$ by noticing that the standard basis vectors[24] $\mathbf{e}_j = [0\ 0\ \cdots\ 0\ 1\ 0\cdots 0]$ are eigenvectors of $\mathbf{G}$. Then $\mathbf{u}_j = \mathbf{e}_j$ and

$$c_j = \frac{\left\langle\mathbf{u}_j, \mathbf{z}^{(0)}\right\rangle}{\|\mathbf{z}^{(0)}\|} = \frac{z_j^{(0)}}{\|\mathbf{z}^{(0)}\|}$$

where $z_j^{(0)}$ is the jth entry of $\mathbf{z}^{(0)}$. As entries of $\mathbf{z}^{(0)}$ are drawn from $\mathcal{N}(0, 1)$, the squared norm of $\mathbf{z}^{(0)}$ follows a Chi-squared distribution with $N - 1$ degrees of freedom. Write $C_{\mathcal{N}(0,1)}(a)$ as the critical value $\mathcal{N}(0, 1)$ and probability a, and $C_{\chi^2}(a, N)$

is the critical value for χ^2_N and probability a. Then, with probability $1 - 2a$,

$$\frac{C_{\mathcal{N}(0,1)}(1-a)}{\sqrt{C_{\chi^2}(1-a, N-1)}} \leq c_i \leq \frac{C_{\mathcal{N}(0,1)}(a)}{\sqrt{C_{\chi^2}(1-a, N-1)}}.$$

Due to the symmetry of $\mathcal{N}(0, 1)$, this is equivalent to

$$|c_i| \leq \frac{C_{\mathcal{N}(0,1)}(a)}{\sqrt{C_{\chi^2}(a, N-1)}}.$$

For $a = 0.99$ and $N = 2000$, we have $C_{\mathcal{N}(0,1)}(a) \approx 2.33$ and $C_{\chi^2}(1-a, N-1) \approx 1854.86$. Then $|c_i| \leq 0.0542$ with probability at least 0.99.

We combine this upper bound on $|c_i|$ with the computed maxima of $\|q_j(x)\|$ over $[0, \lambda_p]$ and the values of $|q_j(\mathbf{G})\mathbf{z}^{(0)}|$ to get upper bounds on $\cos\vartheta\left(\mathbf{u}_i, \mathbf{z}_j\right)$. The results are shown in Table 1, and we compare these with the computed values for $\max_{\mathbf{u}_i \in \mathcal{U}_{\text{noise}}} \cos\vartheta\left(\mathbf{u}_i, \mathbf{z}_j\right)$. The upper bounds on $\cos\vartheta\left(\mathbf{u}_i, \mathbf{z}_j\right)$ are tighter than those produced from Corollary 1. Also, it is clear from Table 1 that the space span $\{\mathbf{z}_2, \mathbf{z}_1\}$ is ρ-free of noise for all $\rho \geq 0.0007$.

Table 1 Computed values for $\max_{0 \leq x \leq \lambda_p} |q_j(x)|$, $\|q_j(\mathbf{G})\mathbf{z}^{(0)}\|$, probabilistic upper bounds on $\cos\vartheta\left(\mathbf{u}_i, z_j\right)$, and computed $\max_{\mathbf{u}_i \in \mathcal{U}_{\text{noise}}} \cos\vartheta\left(\mathbf{u}_i, \mathbf{z}_j\right)$ for the Ritz vectors $\mathbf{z}_j$ from $\mathcal{K}_6\left(\mathbf{G}, \mathbf{z}^{(0)}\right)$. Small values of $\|q_j(\mathbf{G})\mathbf{z}^{(0)}\|$ contribute substantially to large upper bounds on $\cos\vartheta\left(\mathbf{u}_j, \mathbf{z}_i\right)$.

Quantity	$\mathbf{z}_6$	$\mathbf{z}_5$	$\mathbf{z}_4$	$\mathbf{z}_3$	$\mathbf{z}_2$	$\mathbf{z}_1$
$\max_{0 \leq x \leq \theta_6} \lvert q_j(x) \rvert$	8.05×10^{-4}	3.55×10^{-5}	8.32×10^{-6}	4.40×10^{-6}	2.56×10^{-6}	1.28×10^{-6}
$\lVert q_j(\mathbf{G})\mathbf{z}^{(0)} \rVert$	7.4×10^{-4}	7.8×10^{-5}	4.19×10^{-5}	4.68×10^{-5}	2.1×10^{-4}	5.56×10^{-3}
probabilistic max $\cos\vartheta\left(\mathbf{u}_j, \mathbf{z}_i\right)$ c_i bounded with $a = 0.99$	5.87×10^{-2}	2.46×10^{-2}	1.07×10^{-2}	5.09×10^{-3}	6.69×10^{-4}	1.25×10^{-5}
computed max $\cos\vartheta\left(\mathbf{u}_j, \mathbf{z}_i\right)$	4.02×10^{-2}	1.16×10^{-2}	5×10^{-3}	2.37×10^{-3}	3.11×10^{-4}	5.78×10^{-6}

The bounds in Lemma 2 indicate that noise may be due to either a large value of $\max_{0 \leq x \leq \lambda_p} |q_j(x) c_i|$ or due to a relatively small value of $\|q_j(\mathbf{G})\mathbf{z}^{(0)}\|$.

4. Conclusion

We have presented sufficient conditions for a Krylov subspace approximation of the tSVD to be ρ-free of noise. Generally speaking, one can see that minimal Krylov subspace substitutions for the tSVD are doomed to be noisy if n is large enough, even if the start vector is orthogonal to the noise space up to machine precision. However, for moderate n, one can use the sufficient conditions to design a filter to produce a start vector $\mathbf{z}^{(0)}$ that has a small enough $\cos\vartheta\left(\mathbf{u}_i, \mathbf{z}^{(0)}\right)$ for noise space $\mathbf{u}_i$ so that $\mathcal{K}_n\left(\mathbf{G}, \mathbf{z}^{(0)}\right)$ is ρ-free of noise. We are then motivated to find methods to compute good start vectors for minimal Krylov subspaces.

When considering methods to prepare start vectors that satisfy the sufficient conditions presented here, we recall the overarching purpose for minimal Krylov subspace approximations to the tSVD: dramatically smaller compute times. Start vector generation methods that have smaller computational costs are preferable. Start vectors may be implicitly filtered with either Implicitly-Restart Lanczos[25] or Thick-Restart Lanczos[26] when n is small, and Lemma 2 gives a criterion for which Ritz vectors to discard. When n is large, the asymptotic cost of computing Ritz vectors—$O(n^2N)$—may become prohibitive. Implicit start vector filtering may also be less attractive when matrix-vector products scale better than dot products or matrix norms, as is the case for some classes of distributed matrices. Then filtering methods, such as Chebyshev polynomials or approximation of sigmoidal functions[27] may be less expensive. Since $\mathbf{G}$ is a positive semi-definite matrix, simple power iteration on a block vector as in Halko et al.[28] may also be an effective method for preparing a start vector that satisfies our sufficient conditions.

5. References

1. Jolliffe I. Principal component analysis. New York (NY): Springer-Verlag; 2002.

2. Turk M, Pentland A. Eigenfaces for recognition. Journal of Cognitive Neuroscience. 1991;3(1):71–86.

3. Deerwester S, Dumais S, Furnas G, Landauer T, Harshman R. Indexing by latent semantic analysis. Journal of the American Society for Information Science. 1990;41(6):391–407.

4. Hansen PC. Truncated singular value decomposition solutions to discrete ill-posed problems with ill-determined numerical rank. SIAM Journal on Scientific and Statistical Computing. 1990;11(3):503–518.

5. Hansen PC, Sekii T, Shibahashi H. The modified truncated SVD method for regularization in general form. SIAM Journal on Scientific and Statistical Computing. 1992;13(5):1142–1150.

6. Schelter B, Winterhalder M, Timmer J. Handbook of time series analysis: recent theoretical developments and applications. Weinheim (Germany): Wiley-VCH Verlag GmbH & Co. KGaA; 2006.

7. Hyvärinen A, Hurri J, Hoyer P. Natural image statistics: a probabilistic approach to early computational vision. New York (NY): Springer-Verlag; 2009. (Computational Imaging and Vision).

8. Skillicorn D. Understanding complex datasets: data mining with matrix decompositions. New York (NY): Chapman & Hall/CRC; 2007.

9. Liang Y, Lee H, Lim S, Lin W, Lee K, Wu C. Proper orthogonal decomposition and its applications — part I: theory. Journal of Sound and Vibration. 2002;252(3):527–544.

10. Fukunaga K. Introduction to statistical pattern recognition. San Diego (CA): Academic Press; 1990.

11. Blom K, Ruhe A. A Krylov subspace method for information retrieval. SIAM Journal on Matrix Analysis and Applications. 2004;26(2):566–582.

12. Simon H, Zha H. Low-rank matrix approximation using the Lanczos bidiagonalization process with applications. SIAM Journal on Scientific Computing. 2000;21(6):2257–2274.

13. Chen J, Saad Y. Lanczos vectors versus singular vectors for effective dimension reduction. IEEE Transactions on Knowledge and Data Engineering. 2009;21(8):1091–1103.

14. Elden L. Matrix methods in data mining and pattern recognition. Philadelphia (PA): Society for Industrial and Applied Mathematics; 2007.

15. Ren CX, Dai DQ. Bilinear Lanczos components for fast dimensionality reduction and feature extraction. Pattern Recognition. 2010;43(11):3742–3752.

16. Golub G, Luk F, Overton M. A block Lanczos method for computing the singular values and corresponding singular vectors of a matrix. ACM Transactions on Mathematical Software (TOMS). 1981;7(2):149–169.

17. Bai Z. Templates for the solution of algebraic eigenvalue problems. Philadelphia (PA): Society for Industrial Mathematics; 2000.

18. Saad Y. On the rates of convergence of the Lanczos and the block-Lanczos methods. SIAM Journal on Numerical Analysis. 1980;17(5):687–706.

19. Golub G, Van Loan C. Matrix computations. Baltimore (MD): Johns Hopkins University Press; 2013.

20. Hansen PC. The truncated SVD as a method for regularization. BIT Numerical Mathematics. 1987;27(4):534–553.

21. Zhu P, Knyazev A. Angles between subspaces and their tangents. Journal of Numerical Mathematics. 2013;21(4):325–340.

22. Saad Y. Numerical methods for large eigenvalue problems. Manchester (UK): Manchester University Press; 1992.

23. Simon H. Analysis of the symmetric Lanczos algorithm with reorthogonalization methods. Linear Algebra and Its Applications. 1984;61:101–131.

24. Dettman J. Introduction to linear algebra and differential equations. New York (NY): Dover; 1986.

25. Calvetti D, Reichel L, Sorensen D. An implicitly restarted Lanczos method for large symmetric eigenvalue problems. Electronic Transactions on Numerical Analysis. 1994;2(1):1–21.

26. Wu K, Simon H. Thick-restart Lanczos method for large symmetric eigenvalue problems. SIAM Journal on Matrix Analysis and Applications. 2000;22(2):602–616.

27. Kokiopoulou E, Saad Y. Polynomial filtering in latent semantic indexing for information retrieval. Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval; 2004 July 25–29; Sheffield, (UK). New York, (NY): ACM; 2004. p. 104–111.

28. Halko N, Martinsson P, Tropp J. Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions. SIAM Review. 2011;53(2):217–288.

Intentionally left blank.

1 (PDF)	DEFENSE TECHNICAL INFORMATION CTR DTIC OCA
2 (PDF)	DIRECTOR US ARMY RESEARCH LAB RDRL CIO L IMAL HRA MAIL & RECORDS MGMT
1 (PDF)	GOVT PRINTG OFC A MALHORTA
2 (PDF)	DIR USARL RDRL CIH C E CHIN A BREUER

Intentionally left blank.